

## Appendix A: model diagnostics

We used Stan (Carpenter et al., 2017) as the probabilistic language behind the estimation of our Bayesian models in this study, with brms as its R interface (Bürkner, 2017). This language implements the Markov Chain Monte-Carlo (MCMC) algorithm to explore the posterior distribution of the model. Broadly, this algorithm is used to iteratively sample the joint sampling space of the parameters to be estimated in the model, and compute, for each value sampled, its likelihood under some probability distribution previously defined. We run 4 MCMC chains, each 1000 iterations long each. The correct performance of this algorithm is critical to the quality of the statistical evidence to which the outcomes of the model lead.

One way to diagnose the behaviour of MCMC is to inspect whether the different MCMC chains (if more than one) have converged to a similar region of the posterior. The Gelman-Rubin diagnostic ( $\hat{R}$  or R-hat Gelman & Rubin, 1992) provides a measure of chain convergence by comparing the variance within each chain *versus* the variance between each chain. Both are expected to be identical when chains have perfectly converged, so that  $\hat{R} = 1$ . Values lower than 1.01 are recommended, while values higher than 1.05 indicate that chains might have trouble converging and therefore the estimated parameters must be taken with caution. Figure 1 (A) shows the distribution of  $\hat{R}$  values for the coefficients of the fixed effect of our models, which we used for statistical inference. Most values are lower than 1.01, and never higher than 1.05, which provides evidence of successful MCMC convergence.

Another diagnostic of good MCMC converge is the ratio of effective sample size to total sample size ( $N_{eff}/N$ ), which indicates the proportion of samples in the chain that resulted from a non-divergent transition. Values closer to 1 are ideal, as they indicate that all posterior samples from the MCMC were used to estimate the posterior distribution of the parameter. Values larger than 0.1 are recommended. Figure 1 (B) shows the distribution of the effective sample sizes of the coefficients of the fixed effects in our models. Most values are larger than 0.1, although model 0 ( $\mathcal{M}_0$ ) accumulates most effective sample sizes close to 0.1.

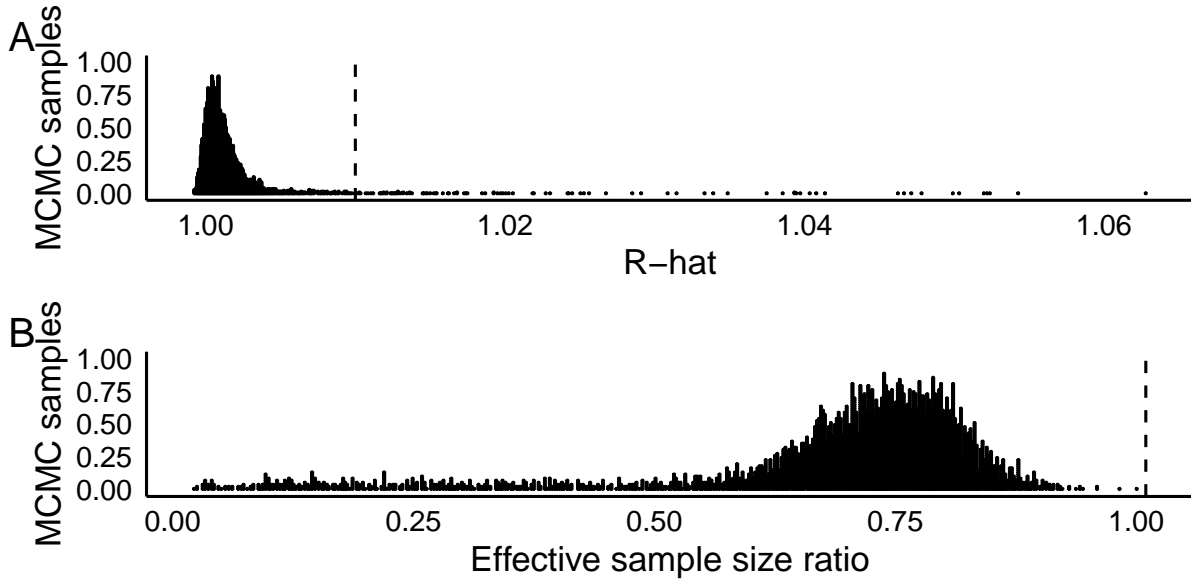


Figure 1: MCMC convergence diagnostic of all parameters in the model. A: distribution of the Gelman-Rubin (R-hat) scores. B: distribution of the ratio of effective sa

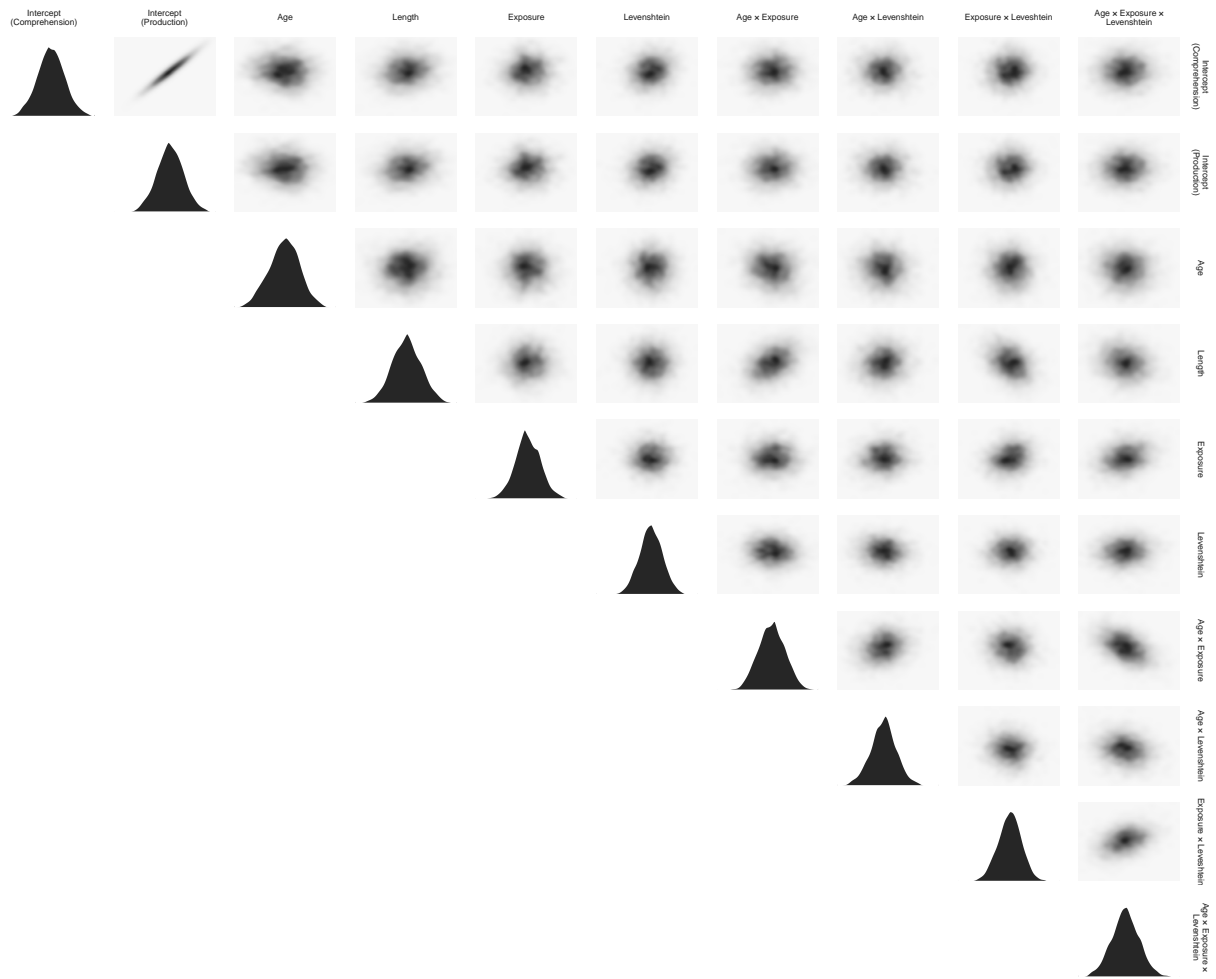
23 **Posterior draws and bi-variate scatterplots**

Figure 2: Marginal distribution and bi-variate scatterplot of posterior samples for the fixed regression coefficients in Model 3.

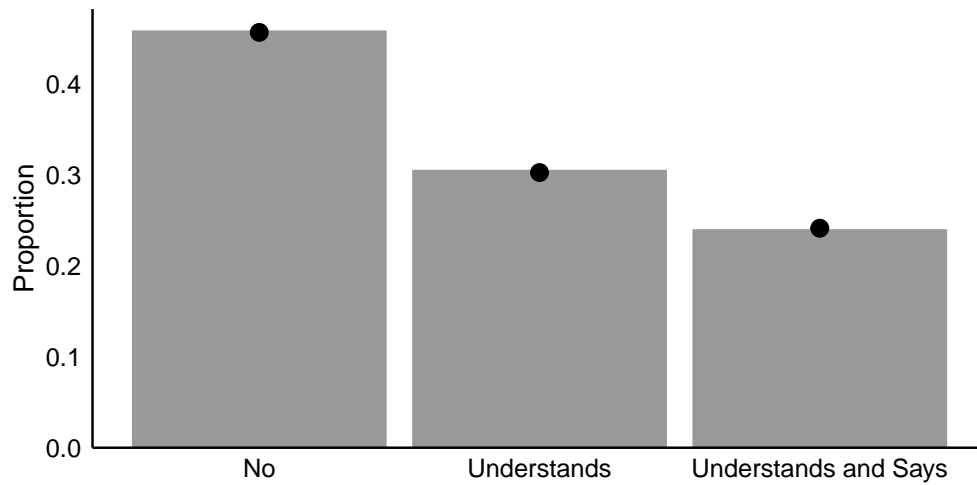
24 **Posterior-predictive checks**

Figure 3: Model posterior predictive checks (PPC). Bars indicate the observed proportion of responses to each category (No, Understands, and Understands and Says). Blue dots and error bars represent the mean proportion of responses simulated from the posterior for each category, and its 95% interval.

Table 1: Posterior distribution of regression coefficients of the model including the *Exposure* composite predictor, and of the model including lexical frequency (*Frequency*) and degree of exposure (*DoE*) separately. Median: median of the posterior distribution in the probability scale. 95% HDI: 95% highest density interval of the distribution.  $p(\text{ROPE})$ : overlap between the 95% HDI and the ROPE, indicating the posterior probability that the true value of the coefficient is equivalent to zero. In the right column, we show a graphical representation of the median and 95% HDI of each coefficient, with a reference dotted line indicating the location of zero.

	$\beta$	95% HDI	$p(\text{ROPE})$
<b>Model: Exposure</b>			
Length (+1 SD, 1.56 phonemes)	0.485	[0.478, 0.491]	.000
Age (+1 SD, 4.86 months)	0.600	[0.588, 0.611]	.000
Exposure (+1 SD, 1.81)	0.558	[0.55, 0.567]	.000
Cognateness (+1 SD, 0.26)	0.515	[0.504, 0.526]	.120
Exposure $\times$ Cognateness	0.486	[0.483, 0.488]	.000
Age $\times$ Exposure	0.518	[0.51, 0.526]	.000
Age $\times$ Cognateness	0.504	[0.5, 0.506]	.974
Age $\times$ Exposure $\times$ Cognateness	0.495	[0.493, 0.498]	.907
<b>Model: Frequency &amp; DoE</b>			
Age (+1 SD, 4.86, months)	0.601	[0.59, 0.612]	.000
Phonemes (+1 SD, 1.56 phonemes)	0.486	[0.479, 0.492]	.000
Frequency (+1 SD, 0.19)	0.527	[0.516, 0.54]	.000
DoE (+1 SD, 0.3)	0.558	[0.549, 0.566]	.000
Cognateness (+1 SD, 0.26)	0.516	[0.506, 0.528]	.026
Age $\times$ DoE	0.518	[0.51, 0.527]	.000
DoE $\times$ Cognateness	0.486	[0.483, 0.488]	.000
Age $\times$ Cognateness	0.504	[0.501, 0.507]	.898
Age $\times$ DoE $\times$ Cognateness	0.495	[0.493, 0.498]	.901

## Appendix B: frequency and language exposure as separate predictors

As a robustness check, we fit a model similar to the one described in the main manuscript, but including lexical frequency and language degree of exposure as separate predictors, instead of the composite measure *Exposure*. Language degree of exposure (*DoE*) was included in interaction with *Age* and *Cognateness*, while lexical frequency (*Frequency*) was included as a main effect. Table 1 shows a comparison between the posterior distribution of the regression coefficients of both models. Overall, results are equivalent.

Table 2: Sample of items included in the BVQ questionnaire and their syllabified SAMPA transcriptions in Catalan and Spanish

Translation	Item	X-SAMPA	Syllables	Item	X-SAMPA	Syllables
cookie	galleta	g@"5E.t@	3	galleta	ga"Le.ta	3
pigeon	colom	ku"lom	2	paloma	pa"lo.ma	3
bottle	biberó	bi.B@"4o	3	biberón	bi.Be"4on	3
book	llibre	"Li.B4@	2	libro	"li.B4o	2
lion	lleó	L@"o	2	león	le"on	2
jacket	jaqueta	Z@"kE.t@	3	chaqueta	tSa"ke.ta	3
smoothie	batut	b@"tut	2	batido	ba"ti.Do	3
octopus	pop	pop	1	pulpo	"pul.po	2
knife	ganivet	g@.ni"BEt	3	cuchillo	ku"tSi.Lo	3
deer	cérvol	"sE4.vol	2	ciervo	"Tje4.Bo	2
steak	filet	fi"IE	2	filete	fi"le.te	3
park	parc	park	1	corralito	ko.ra"li.to	4
airplane	avió	@.Bi"o	3	avión	a"Bjon	2
wheel	roda	"rO.D@	2	rueda	"rwe.Da	2
photo	fotos	"fo.tus	2	fotos	"fo.tos	2

### Appendix C: frequency and language exposure as separate predictors

We define syllable frequency as the rate of appearance of individual syllables in the word-forms included in the Barcelona Vocabulary Questionnaire (BVQ) (Garcia-Castro et al., 2023). Each item corresponds to a Catalan or Spanish word, and has an associated phonological transcription in X-SAMPA format (Wells, 1995). These transcriptions are syllabified. Some examples:

Most Catalan and Spanish words had two syllables, with Spanish words having three and four syllables more often than Catalan words. Less than 1% of the words included in the analyses presented in the main body of the manuscripts had five syllables. No words had more than five syllables (see Figure 4). We extracted lexical frequencies from the English corpora in the CHILDES database (MacWhinney, 2000; Sanchez et al., 2019). Using the Catalan and Spanish corpora was not possible due to the low number of children and tokens included in the corpora.

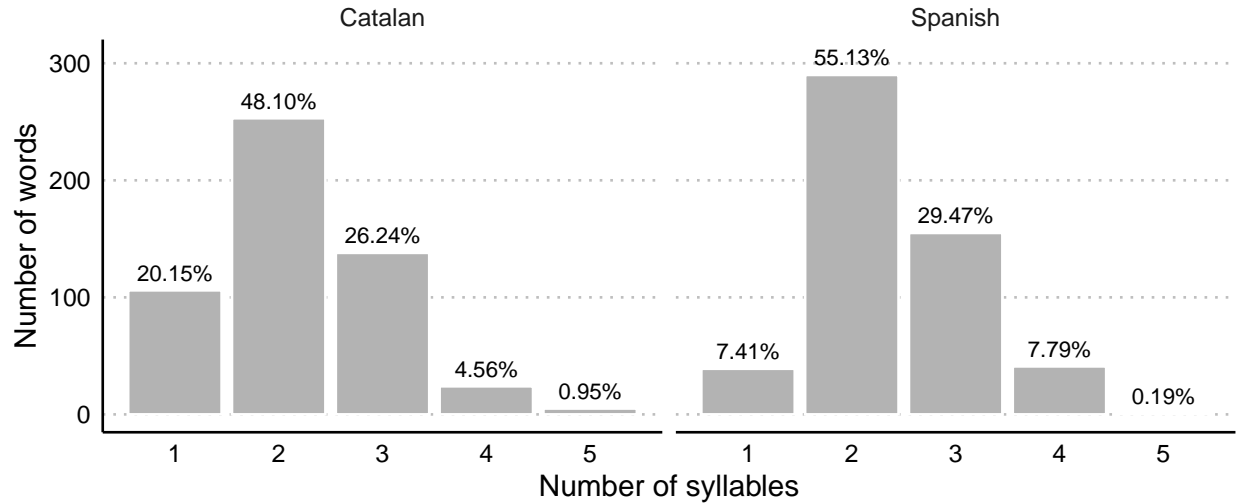


Figure 4: Distribution of the number of syllables in Catalan and Spanish

We now present how syllable frequencies were calculated. Every exposure to a word-form also counts as a exposure to each of the syllables that make up such word. Every time a child hears the word *casa* [house], they are exposed to the syllables *ca* and *sa*. Syllables that appear embedded in words with higher lexical frequency will also be more frequent.

44 To compute the relative frequency of each syllable in Catalan and Spanish (i.e., how many times the syllables appears  
 45 in every million words in Catalan or Spanish speech), we summed the relative lexical frequency in CHILDES of every  
 46 word that contains such syllable in the corresponding language. Figure 5 shows the distribution of frequencies across  
 47 syllables in Catalan and Spanish. In the log10 scale, syllable frequencies in Catalan and Spanish followed a slightly  
 48 asymmetric distribution, with most syllables scoring around 1,000 counts per million, and a longer tail to the right of  
 49 the distribution.

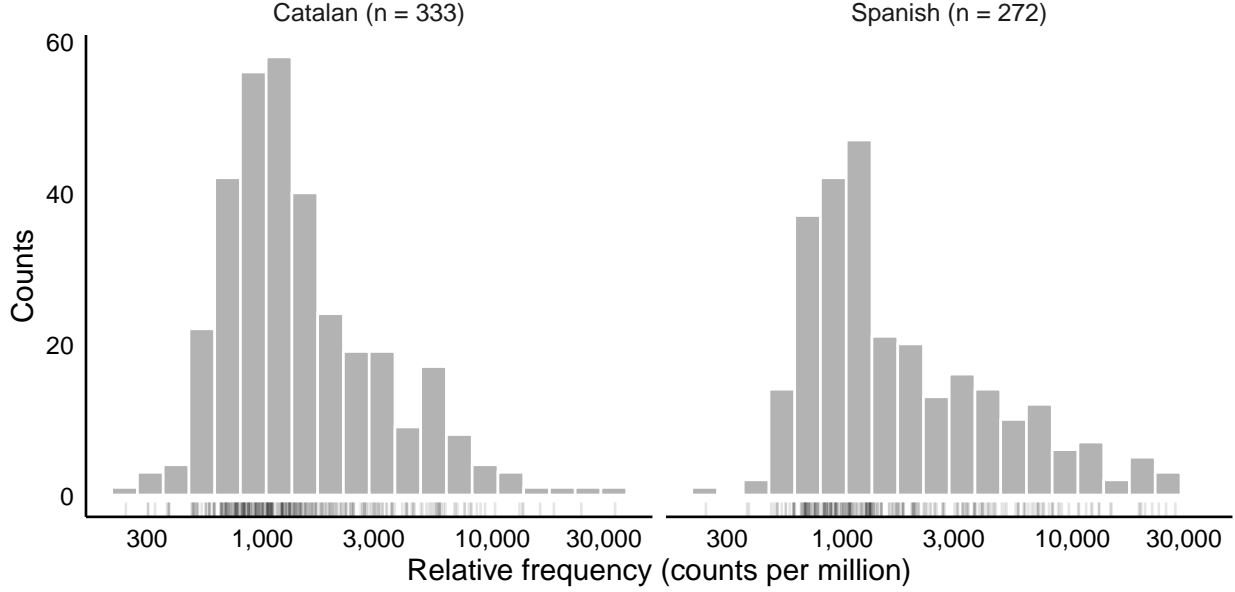


Figure 5: Distribution of apositional syllable frequencies in Spanish and Catalan

50 To estimate the association between word-level syllabic frequency and cognateness, while controlling for the number  
 51 of syllables in the word, as words are expected to necessarily increase the syllabic frequency of the word), we fit  
 52 a multilevel, Bayesian linear regression model with syllabic frequency (the sum of the syllabic frequency of the  
 53 syllables in a word) as response variable, and the main effect of the number of syllables (*Syllables*) and *Cognateness*  
 54 (Levenshtein similarity between a word and its translation equivalent, [Levenshtein, 1966](#)) as predictors. We added  
 55 translation equivalent-level random effects for the intercept and the main effect of *Syllables* (some translation pairs  
 56 had a different number of syllables in each language). We used a Gaussian distribution to model syllabic frequency  
 57 scores after standardising this variable and the predictors. We used a weakly informative prior for all parameters  
 58 involved in the model (see Equation 1 for a formal equation of this model and its prior). We conducted statistical  
 59 inference by evaluating the proportion of the 95% highest density interval (HDI) of the posterior posterior distribution  
 60 of each coefficient that fell into the region of practical equivalence (ROPE, see the main manuscript for a more detailed  
 61 explanation, [Kruschke & Liddell, 2018](#)).

$$\begin{aligned}
 \text{Syllable frequency} &\sim \mathcal{N}(\mu, \sigma) \\
 \mu &= (\beta_0 + u_{0_i}) + (\beta_1 + u_{1_i})\text{Syllables} + \beta_2\text{Cognateness} \\
 \beta_{0-3} &\sim \mathcal{N}(0, 10) \\
 u_{0-1_i} &\sim \mathcal{N}(0, \sigma_{u_i}) \\
 \sigma_y &\sim \text{Exponential}(2) \\
 \begin{pmatrix} u_{0_i} \\ u_{1_i} \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_u\right) \\
 \Sigma_u &= \begin{pmatrix} \sigma_{u_0} & \rho_{u_0} \sigma_{u_0} \sigma_{u_1} \\ \rho_{u_1} \sigma_{u_1} \sigma_{u_0} & \sigma_{u_1} \end{pmatrix} \\
 \sigma_{u_{0-1}} &\sim \mathcal{N}_+(1, 0.1) \\
 \rho_u &\sim \text{LKJcorr}(2)
 \end{aligned} \tag{1}$$

Table 3: Posterior distribution of regression coefficients.

	$\beta$	95% HDI	$p(\text{ROPE})$
Intercept	16.089	[16.02, 16.164]	NA
Syllables (+1 SD, 0.802)	5.642	[5.574, 5.709]	.000
Cognateness (+1 SD, 0.24)	0.008	[-0.057, 0.074]	1.000

We fit this model running 4 sampling chains with 1,000 iterations each. Table 3 shows a summary of the posterior distribution of the fixed effects in the model. As expected, words with more syllables scored higher in syllabic frequency: all posterior draws for the regression coefficient of the main effect of this predictor fell outside the ROPE defined between -0.5 and +0.5 ( $\beta = 5.64$ , 95% HDI = [5.57, 5.71]). Keeping the number of syllables constant, the effect of cognateness was negligible: all of the posterior distributions of this predictor fell within the ROPE, providing evidence that the true value of the increment in syllabic frequency for every increase in cognateness is equivalent to zero ( $\beta = 0.01$ , 95% HDI = [-0.06, 0.07]).

Figure 6 shows the median posterior-predicted syllabic frequencies for words with one to four syllables, for the whole range of cognateness values. Overall, cognate words' syllabic frequency is equivalent to that of non-cognates. This suggests that the cognate facilitation effect in word acquisition reported in the present study is not the result from an association between cognateness and higher syllabic frequencies.

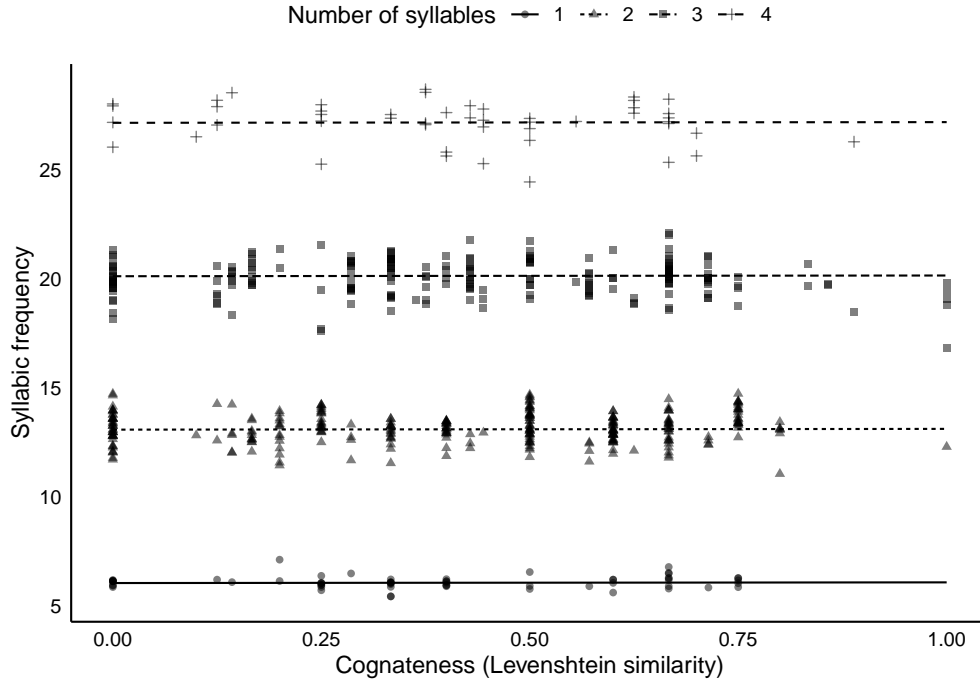


Figure 6: Posterior-predictions of the syllabic frequency model. Thicker lines indicate the median of the posterior predictions, and thinner lines indicate individual posterior predictions.

## References

- Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80, 1–28. <https://doi.org/10.18637/jss.v080.i01>
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan : A probabilistic programming language. *Journal of Statistical Software*, 76(1). <https://doi.org/10.18637/jss.v076.i01>
- Garcia-Castro, G., Ávila-Varela, D. S., & Sebastian-Galles, N. (2023). *Bvq: Barcelona vocabulary questionnaire database and helper functions*. <https://gongcastro.github.io/bvq>
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472. <https://www.jstor.org/stable/2246093>
- Kruschke, J. K., & Liddell, T. M. (2018). The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and planning from a bayesian perspective. *Psychonomic Bulletin & Review*, 25, 178–206. <https://doi.org/10.3758/s13423-016-1221-4>
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics-Doklady*, 10, 707–710.
- MacWhinney, B. (2000). *The CHILDES project: The database* (Vol. 2). Psychology Press.
- Sanchez, A., Meylan, S. C., Braginsky, M., MacDonald, K. E., Yurovsky, D., & Frank, M. C. (2019). Childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior Research Methods*, 51(4), 1928–1941.
- Wells, J. C. (1995). *Computer-coding the IPA: A proposed extension of SAMPA*. 4(28), 1995.